Probability versus Randomness

Hemanshu Kaul, kaul@iit.edu
Department of Applied Math, Illinois Tech

We all warn students about "Gambler's Fallacy" when talking about probability and coin flips, but it's nice to see some real life examples in this BBC article at
http://www.bbc.com/worklife/article/20200217-the-simple-maths-error-that-can-lead-to-bankruptcy.
I particularly found it interesting that higher intelligence might make it more likely to fall victim to this fallacy. I feel that people with a better feeling for mathematics are more likely to confuse the two related notions of "probability" and "randomness". People who like numbers have a strong intuitive idea of randomness that can be misleading especially when confused with probability.

Let me say a bit about what I mean by this. It's mostly an excuse to tell students about a beautiful mathematical idea that shows us how to understand "randomness".

Suppose somebody gives you a few sequences of heads (H) and tails (T) flips of a fair coin (total 30 flips in each sequence; think of 30 as being a very large number ;-)

(0) H H H H H H H H H H H H H H H H H H H H H H H H H H H H H H

(1) H H H H H H H H H H H H H H H H T T T T T T T T T T T T T T

(2) H T H T H T H T H T H T H T H T H T H T H T H T H T H T H T

(3) H T H T T T H H T H H T H T H T H T T T H T H T H T H T H H H

(4) H T T H H H T T T T T T T H H H T H T H H H T H H H T H T H

We know that the probability of each of these sequences actually occurring is the same, $(1/2)^{30}$. And we know that the next outcome in each of these sequences of flips will be H or T with equal probability. But if you have some intuition about numbers, you tend to think that something is unnatural/ non-random about some of these sequences, and you think that unnaturalness has an influence on the next outcome because your intuition tells you that ultimately all such sequences have to look "random". This would lead you to erroneously conclude that the next coin flip in, say, sequence (0) is more likely to be T than H in order to move towards "randomness". (Of course, if this was true then you can make money by betting on this outcome; more about that in the end).

But what does "randomness" really mean? How can we mathematically describe what a "random" sequence of H and T looks like? If the sequences above were given to you as

"data" for a coin flipping experiment, which ones would you believe to be true/ unmanipulated data? Solomonoff and Kolmogorov, independently, in 1960s gave a beautiful answer to this question and laid the foundation of what's called Information or Kolomogorov complexity. Li and Vitanyi's textbook 'Introduction to Kolmogorov Complexity' is a good way to learn about it. Let me give you an intuitive idea about it.

Looking back at the sequences (0), (1), (2), I think most of us would agree that they don't look "random". What about (3) and (4)? Its harder to decide for those.

First, let us try a purely probability or statistics based idea. Look at the pattern of "runs" in each sequence. A run is a contiguous sequence of identical flips - a sequence of uninterrupted repeated H or of uninterrupted repeated T. The current run ends and a new run starts when H follows T or T follows H. Now count the number of runs in each sequence.

In (0) there is exactly 1 run. In (1) there are exactly 2 runs. In (2) there are 29 runs. In (3) there are 23 runs. And in (4) there are 15 runs. We can prove that a "random" sequence of 30 coin flips will on average have around 15 runs (students who have taken probability should be able to prove the exact formula for this in terms of number of flips; try it). This tells us (0), (1), (2), (3) are very unlikely to be random sequences (real/unmanipulated data) as the number of runs are "too far" from 15. While (4) doesn't fail this test, so it could be truly random. "Close" or "far" in this sense can be made precise, in fact this test is more precise the longer a sequence is. It's more difficult to see whether a short sequence is random. However, this test of runs is not enough to characterize randomness. Consider this example of 30 coin flips:
(5) H T H T H T H T H T H T H T H H H H H H H H H H H H H H H
It has exactly 15 runs like (4) but it seems obviously not random ("fake").

So we need something more fundamental to distinguish something like (4) from (5).

Here's a different way of recognizing what's off about these examples: all of them have 'short descriptions', that is they are 'compressible'.

For example (1) can be described as 16 H followed by 14 T. (2) as 14 HT followed by 2H. (5) as 7 HT followed by 16 H. From a programming perspective, these descriptions can be thought of as programs that are much shorter than simply printing the whole sequence of H and T one flip at a time. We say such sequences are compressible (have a much shorter description then their full length). This difference becomes even more startling when you think of examples that are much longer than 30 flips.

Lack of this compressibility means that there are no "patterns" in the truly random sequence of flips data to describe it using memory much less than its length. The Kolmogorov complexity of a sequence of outcomes like our coin flips, is the size of the shortest

computer program needed to generate it. I am ignoring technicalities here (e.g. how to give a precise definition of a computer program and the size of it) but this captures the essence of this idea - so elegant and natural for anyone who has written algorithms or programs. Random sequences don't have any program much shorter than essentially just writing it out step-by-step, one outcome at a time.

Solomonoff and Kolmogorov were able to prove that this property of 'non-compressiblity' is what characterizes 'randomness'.

Let me conclude by going back to the initial discussion where I said that if we could predict (with higher than 1/2 probability) that the next flip would be T then we could make money betting on such an outcome. Say you were playing a simple betting game of "double or nothing" based on a sequence of coin flips. Before each coin flip, based on previous coin flips that you have already seen, you bet a certain amount of your money, from zero to everything, on H and remaining on T. You get the double of your bet amount on the correct outcome and lose all of your bet amount on the incorrect outcome (this is an example of Martingale, some of you might have studied Martingales in your courses). Then, it has been proved that, a truly random sequence of coin flips can be characterized as one where there is no strategy implementable on a computer for making money off this betting game!!!

Another connection that many of you might appreciate is "Expected Kolmogorov Complexity Equals Shannon Entropy"; so yes, Shannon Entropy which measures average information content of the distribution of a random variable is related to Kolmogorov complexity. And Entropy is a notion you might have come across as it is used widely in mathematics and computer science.